Instrumental Variable Bias with Censored Regressors

Roberto Rigobon and Thomas M. Stoker*

March 2005

Abstract

This note examines the inconsistency of instrumental variables (IV) estimators when endogenous regressors are censored. Using a bivariate setting, asymptotic bias is characterized for various types of random censoring and for top-coding and bottom-coding. With independent censoring, expansion bias always occurs, with IV estimates proportionately too large. With dependent censoring, expansion bias typically arises, and we study conditions that assure this with top-coding and bottom-coding. We illustrate these structures graphically for the case with random assignment. Various aspects of estimation with censored regressors are discussed.

1. Introduction

In empirical work, it is common for researchers to use observed variables that are only imperfect reflections of the economic influences or phenomena that they want to capture. When possible, empirical methods are adjusted for such drawbacks in observed data. The most familiar structure of this form is with errors-in-variables, where observed data are assumed to represent a true variable of interest up to random error. When there are other observed indicators of the true variable, they will be used as instruments in consistent estimation. In fact, the approach of using instrumental variables originated because of concerns about errors-in-variables, and now is applied widely to situations of endogeneity, including when important regressors are not observed and must be omitted through necessity.

It is fair to say that instrumental variables (IV) estimation has become the standard method of addressing data problems in empirical work, almost viewed as a cure-all. In some circles, the existence of a valid instrument is the defining criterion of whether empirical results are

*Sloan School of Management, MIT, 50 Memorial Drive, Cambridge, MA 02142 USA; email: rigobon@mit.edu, tstoker@mit.edu. We are grateful to Jerry Hausman and Mitali Das for valuable comments.

of interest at all. When the values of many variables under study result from an unknown simultaneous process, it is arguably necessary to have observations grouped in a way that reflects random assignment, or in other ways reflect a natural experimental design. Understanding the role of randomization in measuring economic effects has been a major conceptual advance in econometrics over the past decade.

However, "cure-all" is too strong a term to apply to IV methods. Data imperfections come in many forms, and many do not admit simple cures with linear IV estimators. A mundane but important example is when a variable is mismeasured proportionally. If that variable is used in level form (namely, not in log-form), it will be difficult to find an instrument to (linearly) separate the true variable from the error. One must model the proportional error and apply instruments in a way that respects the multiplicative structure.

In this paper we study IV estimation with another kind of data imperfection, namely censoring. The simplest form of censoring, as we study here, is where an observed variable equals a true variable of interest for some observations, but equals a constant for other observations. As with ordinary regression analysis, there are methods for adjusting IV estimates when the dependent (left-hand-side) variable is censored.[1] We focus on the implications of censoring a regressor, or an endogenous predictor on the right-hand-side of the equation. We analyze how IV estimates are biased when a censored regressor is used.

There are at least two main sources of censoring of regressors. The first is data mismeasurement, which is sometimes systematic and sometimes not. Observed data may be mismeasured because of bounding, as in top-coding or bottom-coding. For an example of top-coding, household survey data will typically record household income up to a bound of (say) $100,000, with higher income values recorded as "$100,000 or higher."[2] Alternatively, data may be mismeasured at random or missing at random, where a value (say 0) is recorded for each mismeasured or missing value. We could have a combination of these problems – a lower bound of 0 on a variable together with recording of 0 for randomly missing observations.

The second source of censoring is the use of proxies. One may want to represent a certain economic influence in a model, but the best one can do is observe a censored version of that influence. Suppose in a study of consumption, one wants to include the net wealth of each household. Typically, observed wealth is the sum of stock market holdings, housing assets and so forth, but excludes household debt. Observed wealth will then be an imperfect proxy of net financial position, since it excludes the debt component. Observed wealth is also censored (bottom-coded) at 0, since it is never negative. This censoring will cause net financial position

[1] See Blundell and Smith (1989,1994), Hong and Tamer (2003) and Blundell and Powell (2004), among others.

[2] We consider censoring with a single censoring value – here all high incomes are censored and (could be) recorded as the upper bound $100,000. For simplicity, we do not address problems of more general censoring and data coarsening in this paper. For instance, if income values are only reported in broad ranges (such as 0-$25,000, $25,001-50,000, etc.), there is interval censoring. See Manski and Tamer (2002), for instance.

to be very badly proxied for poor households, in particular. For another example, in a study of corporate finance, one may use observed dividends to proxy cash flexibility (absence of cash constraints) Observed dividends are bounded below by 0, and can be viewed as a censored version of the level of cash flexibility.

There has been very little work on the implications of using censored regressors. Rigobon and Stoker (2005) show that using censored regressors implies bias in OLS estimates. They show how different censoring mechanisms can give rise to either attenuation bias (estimates too small) or expansion bias (estimates too big).

In paper, we examine the analogous problem in the context of IV estimation. Our basic setup is a bivariate model with an endogenous regressor, together with a valid instrument. When the regressor is censored, the IV estimator is biased. Mathematically, this bias can be positive or negative. However, we argue that the predominant situation will involve expansion bias, with the IV estimator proportionately too large.

We begin by illustrating the basic structure of IV bias in a situation where there is random assignment to two groups. We consider various types of censoring, from independent censoring to dependent censoring such as top-coding and bottom-coding. After that, we show a general characterization of the IV bias, and then apply it to the different types of censoring.

Independent censoring is particularly interesting. This always produces expansion bias. The expression of the bias depends only on the percentage of data censored, not on the distribution of the endogenous regressor or instrument, or on the censoring value. In contrast, with OLS estimates, independent censoring produces attenuation bias (or in a special case, zero bias); see Rigobon and Stoker (2005).

With censoring that is correlated with the regressor and/or the instrument, expansion bias arises unless the censored data have a dramatically different correlation structure to the uncensored data. For top-coding and bottom-coding, we characterize this situation precisely, and indicate a condition of "mean-monotonicity" that assures expansion bias occurs. In all cases, we view expansion bias as the "typical" outcome.

After presentation of the various results on IV bias, we discuss some aspects of estimation. In particular, we consider estimation with complete cases only, estimation with censored instruments, and methods for estimation with the full data sample.

2. Instrumental Variables Estimation with Censored Regressors

2.1. Basic Framework and Bias

We can get several insights from studying instrumental variables estimation in a bivariate linear model. Suppose that the true model is

$$y_i = \alpha + \beta x_i + \varepsilon_i \quad i = 1, ..., n \tag{1}$$

where x_i is the (uncensored) regressor, ε_i is the disturbance and z_i denotes a valid instrument for x_i. In particular, we assume that $\{(x_i, z_i, \varepsilon_i) \mid i = 1, ..., n\}$ is an i.i.d. random sample from a distribution with finite second moments, with $E(\varepsilon) = 0$, $Cov(z, x) \neq 0$ and $Cov(z, \varepsilon) = 0$. The IV estimator $\hat{\gamma}$, that uses z_i to instrument x_i, is consistent for β; as in

$$\hat{\gamma} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i - \bar{x})} \to \frac{Cov(z, y)}{Cov(z, x)} = \beta \tag{2}$$

Suppose that we do not observe the variable x_i, but rather a censored version of it. Let d_i indicate a general censoring process, where $d_i = 1$ indicates that the observed regressor is censored and set to the value ξ, and $d_i = 0$ otherwise. That is, we observe x_i^{cen}, the censored version of x_i, where

$$x_i^{cen} = (1 - d_i) x_i + d_i \xi. \tag{3}$$

At this point d_i is not restricted. d_i can be purely random, possibly correlated with x_i, z_i, or ε_i, or determined more specifically as in top coding ($d_i = 1[x_i > \xi]$) or bottom coding ($d_i = 1[x_i < \xi]$). We denote the probability of censoring as $p = \Pr\{d = 1\} \neq 0$.

What happens if we use x_i^{cen} instead of x_i in estimation? We compute the IV estimator $\hat{d}$ that uses z_i to instrument x_i^{cen}, with

$$\hat{c} = \frac{\sum_{i=1}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1}^{n} (z_i - \bar{z})(x_i^{cen} - \bar{x}^{cen})} \to \frac{Cov(z, y)}{Cov(z, x^{cen})} \tag{4}$$

where we have assumed $Cov(z, x^{cen}) \neq 0$, since otherwise $\hat{c}$ has no probability limit. From (2), we can express the asymptotic bias as

$$\text{plim } \hat{c} = \frac{Cov(z, x)}{Cov(z, x^{cen})} \cdot \beta. \tag{5}$$

We develop the IV bias further by noting that $x_i = x_i^{cen} + x_i^{o}$, where

$$x_i^{o} = d_i (x_i - \xi) \tag{6}$$

is the part of x lost by censoring. As such, we have

$$Cov\,(z,x) = Cov\,(z,x^{cen}) + Cov\,(z,x^{o}) \tag{7}$$

and we write the IV bias in proportional form as

$$\text{plim}\ \hat{c} = (1+\Lambda)\cdot\beta, \quad \text{with} \quad \Lambda = \frac{Cov\,(z,x^{o})}{Cov\,(z,x^{cen})}. \tag{8}$$

There is no bias if $\Lambda = 0$. There is *attenuation bias* if $-1 < \Lambda < 0$, and there is *expansion bias* if $\Lambda > 0$.

We now turn to an analysis of the bias when the instrument represents random assignment. This will motivate the more general results we give later.

2.2. Random Assignment

We can get a lot of intuition from the case of random assignment with two groups. Assume z is a binary instrument indicating groups 0 and 1, and denote the probability of being in group 1 as $q = \Pr\{z = 1\} \neq 0$. Here $Cov(z,x) \neq 0$ implies

$$E\,(x \mid z = 1) - E\,(x \mid z = 0) \neq 0, \tag{9}$$

or that the grouping is associated with a shift in the mean of x. $Cov\,(z,\varepsilon) = 0$ implies

$$E\,(\varepsilon \mid z = 1) = E\,(\varepsilon \mid z = 0) = 0, \tag{10}$$

or that there is no shift in the mean of ε associated with the grouping.

The IV estimator $\hat{\gamma}$ of (2) is the group-difference estimator of Wald (1940), namely

$$\hat{\gamma} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1 - \bar{x}_0} \tag{11}$$

where $\bar{y}_1 = \sum_{i=1}^{n} z_i y_i / \sum_{i=1}^{n} z_i$ is the average of y_i for group 1, $\bar{y}_0 = \sum_{i=1}^{n} (1-z_i)\, y_i / \sum_{i=1}^{n} (1-z_i)$, is the average of y_i for group 0, etc. Obviously

$$\text{plim}\ \hat{\gamma} = \frac{E\,(y \mid z = 1) - E\,(y \mid z = 0)}{E\,(x \mid z = 1) - E\,(x \mid z = 0)} = \beta. \tag{12}$$

When we use the censored regressor x_i^{cen} instead of x_i, the IV estimator is

$$\hat{c} = \frac{\bar{y}_1 - \bar{y}_0}{\bar{x}_1^{cen} - \bar{x}_0^{cen}} \tag{13}$$

where $\bar{x}_1^{cen}$, $\bar{x}_0^{cen}$ are the censored group averages, and we have

$$\text{plim } \hat{c} = \frac{E(y \mid z=1) - E(y \mid z=0)}{E(x^{cen} \mid z=1) - E(x^{cen} \mid z=0)} = (1+\Lambda) \cdot \beta. \tag{14}$$

The proportional IV bias term is

$$\Lambda = \frac{E(x^o \mid z=1) - E(x^o \mid z=0)}{E(x^{cen} \mid z=1) - E(x^{cen} \mid z=0)} \tag{15}$$

where again $x_i^o = d_i(x_i - \xi) = x_i - x_i^{cen}$ is the part of x_i omitted by the censoring, and we have

$$\begin{aligned} E(x \mid z=1) - E(x \mid z=0) &= E(x^{cen} \mid z=1) - E(x^{cen} \mid z=0) \\ &\quad + E(x^o \mid z=1) - E(x^o \mid z=0). \end{aligned} \tag{16}$$

The size and sign of Λ is determined by how the censoring operates on the two different assignment groups.

The IV bias will be zero only in very unusual circumstances where the censoring doesn't remove any information of value for identifying β. In particular, the IV bias is zero only when

$$E(x^o \mid z=1) - E(x^o \mid z=0) = 0. \tag{17}$$

The mean shift (9) that identifies β occurs entirely in the mean of the censored regressor x^{cen}, with no help from the omitted x^o.

When the data that is censored has broadly similar structure to the uncensored data, there is expansion bias. From (15) and (16), it is clear that $\Lambda > 0$ when the mean shifts $E(x^o \mid z=1) - E(x^o \mid z=0)$ and $E(x^{cen} \mid z=1) - E(x^{cen} \mid z=0)$ are of the same sign. Alternatively, there is attenuation bias ($\Lambda < 0$) only when the mean shifts are of opposite signs.

A most interesting case of expansion bias is when censoring is purely random, where d is statistically independent of z and x. Here we can solve explicitly for the bias. We have

$$\begin{aligned} E(x^o \mid z=1) - E(x^o \mid z=0) &= E(d) \cdot [E(x - \xi \mid z=1) - E(x - \xi \mid z=0)] \\ &= p \cdot [E(x \mid z=1) - E(x \mid z=0)] \end{aligned} \tag{18}$$

From (15) and (16), the proportional IV bias is

$$\Lambda = \frac{p}{1-p} > 0. \tag{19}$$

It is useful to note that the bias in the IV estimator does not depend on the distribution of x, the assignment grouping or the censoring value ξ. It depends only on p, the percentage censored. The bias is nonlinear and strictly convex in p, with 20% censoring associated with 25% proportional bias, and 50% censoring associated with 100% proportional bias. Later we will show that (19) is valid under much more general conditions.

Figures 1 and 2 illustrate IV bias with random censoring.[3] Figure 1 shows the uncensored data, including the z value grouping. There is substantial overlap between the groups. Also illustrated are the group means and the uncensored IV (group difference) estimator (11). Figure 2 shows what happens with 30% random censoring. The censored data are shown as small circles, with the censoring value $\xi = 4$. The IV fit is clearly steeper, which illustrates the positive IV bias.[4] Mechanically, the within-group means of x are both shifted toward ξ but the within-group means of y are unchanged, tilting the estimator (13) versus (11). It is worth mentioning that the (slope) bias does not depend on the specific censoring value ξ, and the same tilting occurs whether $\xi = 0$ or $\xi = 6$. We verified this but do not include an additional figure.

For a different example with random assignment, consider censoring via top-coding. Here, all values of x above a limit ξ are set to ξ, with the censoring indicator

$$d_i = 1\left[x_i > \xi\right]. \tag{20}$$

Figure 3 illustrates expansion bias with top-coding. The same data are used here as in Figures 1 and 2, and now censoring occurs for values greater than $\xi = 6$. Clearly, much more censoring occurs for observations with $z = 1$ than for those with $z = 0$. Because of top-coding, the mean of the x values for $z = 1$ is shifted more to the left than than the mean of x values for the $z = 0$ group, which gives the positive IV bias.[5]

We would expect expansion bias to be the typical case with top-coding, but it not assured. If the $z = 0$ group had much wider dispersion than the $z = 1$ group, then top-coding could involve censoring more of the right tail of the $z = 0$ group than that of the $z = 1$ group. If the mean of x values for the $z = 0$ group is shifted more than the mean for $z = 1$, then attenuation

[3] The specification is as follows. The true model is $y_i = 2 + .5x_i + \varepsilon_i$. The probability of $z_i = 1$ is $q = .6$. For $z_i = 0$, (x_i, ε_i) is joint normal with mean $(2, 0)$, where the standard error of x_i is 2, the standard error of ε_i is 1 and the correlation between x_i and ε_i is $-.5$. For $z_i = 1$, we have (x_i, ε_i) joint normal with mean $(6, 0)$ and the same variance parameters.

[4] The value of the uncensored estimator is $\hat{\delta} = .46$ and the censored estimator is $\hat{d} = .64$. The actual fraction of data censored is .28.

[5] Here we have $\hat{\delta} = .46$ as before, and $\hat{d} = .59$ with top-coding. The fraction of data censored is .34.

bias would result.[6]

To learn more, we can apply a little probability arithmetic to get the formula

$$\begin{aligned} E\left(x^o \mid z=1\right) - E\left(x^o \mid z=0\right) &= p \cdot \left[E\left(x \mid d=1, z=1\right) - E\left(x \mid d=1, z=0\right)\right] \\ &\quad + \left(\frac{Cov\left(z,d\right)}{q}\right) \cdot \left[E\left(x \mid d=1, z=1\right) - \xi\right] \\ &\quad + \left(\frac{Cov\left(z,d\right)}{1-q}\right) \cdot \left[E\left(x \mid d=1, z=0\right) - \xi\right]. \end{aligned} \tag{21}$$

for random assignment with a general censoring indicator d. Applying (21) to top-coding shows that the IV bias depends on the censoring percentage p, the distributions of the x's within the groups and the relative locations of the group distributions. If the mean of the observations censored from the $z = 1$ group is at least as large as the mean from the $z = 0$ group, and at least as much censoring occurs in the $z = 1$ group as the $z = 0$ group ($Cov\left(z,d\right) \geq 0$), then all the RHS terms of (21) are positive, and expansion bias ($\Lambda > 0$) arises in the IV estimator. Attenuation bias requires a larger mean for the censored observations for the $z = 0$ group, more censoring for the $z = 0$ group, or both.[7]

The formula (21) also gives us some more primitive conditions for the absence of IV bias. In particular, we will have zero IV bias if the mean of the observations that are censored is the same in both groups:

$$E\left(x \mid d=1, z=1\right) = E\left(x \mid d=1, z=0\right) \tag{22}$$

and

$$\text{either} \qquad Cov\left(z,d\right) = 0 \quad \text{or} \quad \xi = E\left(x \mid d=1, z=1\right)$$

There is more involved than just the mean of the censored observations; either the instrument is uncorrelated with the censoring or the censoring value is exactly equal to the mean.[8] Again, it seems difficult to devise censoring mechanisms that would obey these conditions without assuming specific formulas for the underlying distributions.

[6] With the mixture of normals design of the figures, one could solve for how much the standard deviation of x conditional on $z = 0$ has to increase relative to that of $z = 1$ to generate attenuation bias. We note here some casual observations. Quadrupling the standard error of x to 8 for $z = 0$ produces attenuation bias as the rule. Tripling to 6 produces simulated results with each kind of bias: some with attenuation bias, some with small bias and some with expansion bias.

[7] By symmetry in the logic, all the same remarks apply to censoring with bottom-coding.

[8] In OLS estimation, zero bias occurs if the censoring value is set to the mean of the data lost to censoring. See Rigobon and Stoker (2005).

2.3. Some General Results

We now return to the setting of a general instrument z. The IV bias is given in (8), repeated here as

$$\text{plim } \hat{c} = (1 + \Lambda) \cdot \beta, \quad \text{with} \quad \Lambda = \frac{Cov\,(z, x^o)}{Cov\,(z, x^{cen})}. \tag{23}$$

We give a general characterization of IV bias in the next section, and follow that with results on specific censoring mechanisms.

2.3.1. Preliminaries

We begin by making analogous remarks on bias to those we made above, in terms of the signs and relative values of $Cov\,(z, x^o)$ and $Cov\,(z, x^{cen})$. Zero bias involves the remarkably strong restriction that $Cov\,(z, x^o) = 0$, or that the instrument z_i happens to be correlated with the observed x_i^{cen} but not correlated with x_i^o, the part of x_i that is lost in the censoring.[9] Alternatively, if $Cov\,(z, x^{cen})$ and $Cov\,(z, x^o)$ are of the same sign, then there is expansion bias, with $\Lambda > 0$. If $Cov\,(z, x^{cen})$ and $Cov\,(z, x^o)$ are of opposite signs, then $\Lambda < 0$. This includes attenuation bias with $-1 < \Lambda < 0$, but one cannot rule out $\Lambda < -1$.[10]

Let Δ_z and Δ_x denote the following mean differences between censored and uncensored observations:

$$\Delta_z \equiv E\,(z|d = 1) - E\,(z|d = 0) = \frac{1}{p\,(1-p)} \cdot Cov(z, d), \tag{24}$$

$$\Delta_x \equiv E\,(x|d = 1) - E\,(x|d = 0) = \frac{1}{p\,(1-p)} \cdot Cov(x, d). \tag{25}$$

Some general calculations on IV bias are given as Lemma 1 and Proposition 2. Notes on the proofs of these results are presented in Section 2.3.5.

Lemma 1. *We have*

$$Cov\,(z, x^{cen}) = (1-p)\,Cov\,(z, x|d = 0) + p\,(1-p)\,\Delta_z\,[\xi - E\,(x|d = 0)]\,, \tag{26}$$

$$Cov\,(z, x^o) = pCov\,(z, x|d = 1) + p\,(1-p)\,\Delta_z\,[E\,(x|d = 1) - \xi] \tag{27}$$

[9] Another way of making this obvious is to write the original model (1) in terms of x_i^{cen} instead of x_i. The resulting disturbance is $u = \beta x_i^o + \varepsilon_i$, so that z_i is a valid instrument for x_i^{cen} only when $Cov\,(z, x^o) = 0$.

[10] $\Lambda < -1$ occurs if $Cov(z, x)$ and $Cov(z, x^{cen})$ are of opposite signs, so that censoring radically alters the relationship between z and x. For instance, this will occur if $Cov\,(z, x^{cen})$ is negative and $Cov(z, x^o)$ is positive and larger in absolute value.

and

$$Cov\left(z,x\right)=pCov\left(z,x|d=1\right)+\left(1-p\right)Cov\left(z,x|d=0\right)+p\left(1-p\right)\Delta_z\Delta_x \tag{28}$$

Proposition 2. *The proportional IV bias,* $\Lambda=Cov\left(z,x^o\right)/Cov\left(z,x^{cen}\right)$*, is*

$$\Lambda=\frac{p}{1-p}\frac{Cov\left(z,x|d=1\right)+\left(1-p\right)\Delta_z\left[E\left(x|d=1\right)-\xi\right]}{Cov\left(z,x|d=0\right)+p\Delta_z\left[\xi-E\left(x|d=0\right)\right]} \tag{29}$$

The expression (29) shows that IV bias depends on the distribution of (z,x) for censored and uncensored data through conditional means and covariances, but also on the value ξ that the data is censored to. This mirrors the same point stressed by Rigobon and Stoker (2005) for bias in OLS regression with censored regressors.

We now specialize this formulation to different types of censoring mechanisms. We list bias results for various special cases as a series of Corollaries. Again, notes on their proofs are collected Section 2.3.5

2.3.2. Random Censoring

We now examine the situation of independent censoring as discussed before with random assignment. There must be nonzero expansion bias in this case, and in fact, the bias formula (19) holds quite generally. In particular, we have the following corollary, which includes independence under much weaker correlation assumptions:

Corollary 3. ***Completely Uncorrelated Censoring.*** *If* $Cov(z,d)=0$*,* $Cov(x,d)=0$ *and* $Cov(zx,d)=0$*, then*

$$\Lambda=\frac{p}{1-p} \tag{30}$$

with $\Lambda>0$ *always.*

The conditions are equivalent to assuming that z, x and zx are mean independent of d, and clearly include statistical independence of (z,x) and d. The bias is always positive, varies directly with the amount of censoring p, but is not affected by the distribution of (z,x) or the censoring point ξ.

This result is in strong contrast to the bias in OLS estimators shown in Rigobon and Stoker (2005). Random censoring implies attenuation bias in OLS coefficient estimators, or zero bias in the special case of censoring to the regressor mean, or $\xi=E\left(x\right)$. With the IV estimator,

expansion bias is always implied, regardless of the censoring value.[11]

The role of the distribution and of the censoring point are clarified by some results on 'partial' lack of correlation. Suppose first that we have a situation where only the instrument z is uncorrelated with the censoring. This modifies the IV bias, as in

Corollary 4. ***Censoring Uncorrelated with Instrument*** z. *If* $Cov(z,d)=0$, *then*

$$\Lambda = \frac{p}{1-p}\frac{Cov\left(z,x|d=1\right)}{Cov\left(z,x|d=0\right)} \tag{31}$$

The bias under independence is scaled by the ratio of covariances of z and x over censored and uncensored values. Expansion bias results if the covariances are of the same sign. If the covariance between z and x is much stronger in the uncensored data, then the IV bias will be smaller than with independence.[12]

This scaling, as well as another modification, occurs in the situation where only the regressor is uncorrelated with the censoring, as in

Corollary 5. ***Censoring Uncorrelated with Predictor*** x. *If* $Cov(x,d)=0$, *the*

$$\Lambda = \frac{p}{1-p}\frac{Cov\left(z,x|d=1\right)+\left(1-p\right)\varphi}{Cov\left(z,x|d=0\right)-p\varphi} \tag{32}$$

where $\varphi \equiv \Delta_z \cdot \left[E\left(x\right)-\xi\right]$.

Correlation of the censoring indicator d with the instrument z induces a shift φ in the numerator and denominator of the proportional bias. The shift depends on the censoring value ξ. If z is correlated with d, the shift φ vanishes only when the censoring value is the mean $\xi = E\left(x\right)$. Thus, even if x values are censored at random, IV bias exists, that varies with the censoring point ξ as well the behavior of the instrument over censored and uncensored values.

2.3.3. Top-coding, Bottom-coding and Mean-Monotonicity

We now consider cases of censoring with bounds, where censoring is correlated with the regressor by construction. We first consider top-coding, where values of x are censored above a threshold

[11] While not related to censoring, Black, Berger and Scott (2000) have a similar finding for a specific model of errors-in-variables, where the bias in OLS coefficients can be in the opposite direction of the bias in IV coefficient estimates.

[12] If is worth noting that zero correlation between z and d can include situations where d is correlated with the disturbance ε.

ξ. That is, we have

$$d_i = 1\left[x_i > \xi\right] \tag{33}$$

with

$$x_i^{cen} = x_i \cdot 1\left[x_i \leq \xi\right] + \xi \cdot 1\left[\xi < x_i\right] \tag{34}$$

We set $Cov(z, x) > 0$ without loss of generality.

The intuition discussed in the section on random assignment is enticing here as well. That is, with top-coding, d is positively correlated with x, and x is positively correlated with z, so it seems natural that z and d will be positively correlated and expansion bias in the IV estimate will result. We expect this as the typical case, but unfortunately it is possible for attenuation bias to result, if there is a substantial difference between the distribution of censored and uncensored values. The following corollary gives a fairly general characterization of this:

Corollary 6. *Top-coding.* *Suppose x is top-coded at ξ, $Cov\left(z, x^{cen}\right) > 0$ and $Cov\left(z, d\right) > 0$. Then*

$$\Lambda \leq 0$$

if and only if

$$Cov\left(z, x|d=1\right) \leq -\left(1-p\right) \cdot \Delta_z \cdot \left[E\left(x|d=1\right) - \xi\right] \tag{35}$$

Clearly, $\Lambda > 0$ if $Cov\left(z, x|d=1\right) \geq 0$.

What Corollary 6 says is that there is expansion bias ($\Lambda > 0$) unless (35) holds.[13] We have that both $\Delta_z > 0$ and $\left[E\left(x|d=1\right) - \xi\right] > 0$, so that the right-hand side of (35) is negative. Therefore, the within-covariance of z and x for high x values has to be sufficiently negative to eliminate an upward expansion bias in the IV coefficient. That is, the correlation structure of (z, x) for the high x values has to be radically different, or essentially the opposite, of the correlation structure for the low x values. To the extent that this is unusual, the typical impact of top-coding on IV estimates will be expansion bias.

It would be desirable to discover some primitive conditions that are closely associated with IV expansion bias when there is top-coding. Unfortunately, all the primitive conditions that the authors have discovered are much stronger than the tenets of Corollary 6. Of those, there is one set of conditions worth mentioning, which essentially assures the positive relation between the regressor, instrument and censoring discussed above. This is where z is **mean-monotonic** in x; namely

$$E\left(z|x=x_1\right) \geq E\left(z|x=x_0\right) \text{ for any values } x_1 \geq x_0 \tag{36}$$

This condition guarantees all the conditions of Corollary 6, as summarized in

[13] The condition that $Cov\left(z, x^{cen}\right) > 0$ just assumes that the observed predictor is correlated with the same sign as the uncensored variable.

Corollary 7. ***Top-Coding with Mean-Monotonicity of*** z ***in*** x. *If* x *is top-coded at* ξ *and condition (36) holds, then* $\Lambda > 0$.

Mean-monotonicity gives a uniform structure to the covariance of z and x over ranges of x values. It is not a counterintuitive condition, at least for principal examples. For instance, if (z, x) is joint normally distributed (with nonnegative correlation), then clearly z is mean-monotonic in x.

Suppose we now consider bottom-coding to a lower bound, say ξ. Here we have

$$d_i = 1\left[x_i < \xi\right] \tag{37}$$

with

$$x_i^{cen} = x_i \cdot 1\left[x_i \geq \xi\right] + \xi \cdot 1\left[\xi > x_i\right] \tag{38}$$

and we continue to assume that $Cov(z, x) > 0$.

A moment's reflection indicates that there is no real difference in the structure of IV biases between the situations of bottom-coding and of top-coding. In particular, we summarize this by stating the results for bottom-coding that are analogous to those we just covered.

Corollary 8. ***Bottom-coding.*** *If* x *is bottom-coded at* ξ, $Cov\left(z, x^{cen}\right) > 0$ *and* z *is negatively correlated with* d, *then* $\Lambda \leq 0$ *if and only if (35) holds. Clearly,* $\Lambda > 0$ *if* $Cov\left(z, x | d = 1\right) \geq 0$.

Corollary 9. ***Bottom-Coding with Mean-Monotonicity of*** z ***in*** x. *If* x *is bottom-coded at* ξ *and condition (36) holds, then* $\Lambda > 0$.

Corollary 8 says the same thing for bottom-coding as Corollary 6 did for top-coding: bottom-coding implies expansion bias ($\Lambda > 0$) unless the within-covariance for low x values is sufficiently negative.[14] Again, the correlation structure of (z, x) for the low x values has to be radically different from that of uncensored high x values for IV estimators to not exhibit expansion bias with bottom-coding.[15] That possibility is eliminated with mean-monotonicity of z in x, as echoed by Corollary 9.

The corollaries of the last two section have a common practical lesson. When the regressor is censored, IV estimators will typically exhibit expansion bias. Up to sampling error, estimated effects are too large. For this not to occur, one would have to have a radically different structure between the instrument and the regressor in the observed (censored) data and in the data that has been lost by the censoring.

[14] Here we have $\Delta_z < 0$ and $\left[E\left(x | d = 1\right) - \xi\right] < 0$.

[15] As Rigobon and Stoker (2005) pointed out for OLS estimators, if the data is both top-coded and bottom-coded, the IV estimators will exhibit asymptotic bias terms from each types of censoring. In the typical case of expansion bias, there will be greater IV bias both top-coding and bottom-coding than with only one.

2.3.4. Relation to OLS Bias with a Censored Dependent Variable

Every student of econometrics is familiar with the picture showing that the OLS estimator of a linear model coefficient is downward biased when the dependent variable is censored by bottom-coding (or top-coding).[16] It is useful to note that expansion bias in IV estimators is equivalent to downward bias of OLS in the first stage regression due to censoring of the dependent variable x.

Suppose again that $Cov(z, x) > 0$. We have expansion bias, $\Lambda > 0$, if and only if $Cov(z, x^{cen}) < Cov(z, x)$. This is the same as

$$\frac{Cov(z, x^{cen})}{Var(z)} < \frac{Cov(z, x)}{Var(z)} \tag{39}$$

The right-hand-side of (39) is the limit of the OLS coefficient of the first-stage regression, of x_i regressed on z_i and a constant. The left-hand-side is the limit of the OLS coefficient of the first-stage regression using the censored variable x_i^{cen}.

Thus, the conditions for expansion bias $\Lambda > 0$ are exactly the same as the conditions for downward bias arising in OLS coefficients when the dependent variable is censored, where 'bias' here is defined generally as the difference in the OLS estimator limits with and without censoring. In other words, in examining IV expansion bias we have been equivalently characterizing the conditions that give downward bias in OLS coefficients with dependent variable censoring, doing so without making any assumptions about the form of a "model" between x and z. One immediate conclusion is that the OLS coefficient will be downward biased with top- or bottom-coding when z is mean-monotonic in x.

2.3.5. Notes on Proofs

Lemma 1 and Proposition 2 are direct calculations. The formulae (26) and (27) follow from straightforward arithmetic. For instance, for (26), we have

$$\begin{aligned} Cov(z, x^{cen}) &= Cov(z, (1-d)x) + Cov(z, d) \cdot \xi \\ &= E((1-d)xz) - E(z) \cdot E((1-d)x) + p(1-p)\Delta_z \xi \end{aligned}$$

Write out all of the expectations in terms of expectations conditional on $d = 1$, and simplify to get (26). Equation (27) follows analogously. Finally, equation (28) follows from (7) but also just represents the standard within-between covariance breakdown. Finally, (29) follows from (8).

[16] See Davidson and MacKinnon (2004), Figure 11.3, page 482, among many others.

Corollary 3 follows from noting that three things. First, $Cov(z,d)=0$ implies $\Delta_z=0$ (see (24)). Second, $Cov(zx,d)=0$ and $Cov(x,d)=0$ imply that $E(zx|d=1)=E(zx|d=0)$ and $E(x|d=1)=E(x|d=0)$, so that $Cov(z,x|d=1)=Cov(z,x|d=0)=Cov(z,x)$. Finally, $Cov(z,x|d=0)\neq 0$ since $Cov(z,x)\neq 0$.

Corollary 4, follows from $\Delta_z=0$. Likewise, Corollary 5 is immediate from $E(x|d=1)=E(x|d=0)=E(x)$.

Corollary 6 follows from (29). The final statement is true since $\Delta_z>0$ and top-coding implies $E(x|d=1)-\xi>0$.

Corollary 7 follows from three steps. First, note that $Cov(z,x)\neq 0$ implies that $\zeta(x_0)\equiv E(z|x=x_0)$ is not a constant function. Second,

$$\begin{aligned} Cov(z,d) &= p(1-p)\left[E(z|d=1)-E(z|d=0)\right] \\ &= p(1-p)\left[E(\zeta(x_1)|x_1>\xi)-E(\zeta(x_0)|x_0\leq\xi)\right]>0 \end{aligned}$$

by mean-monotonicity. Finally,

$$\begin{aligned} Cov(z,x|d=1) &= E(zx|d=1)-E(z|d=1)E(x|d=1) \\ &= E\left(z\cdot[x-E(x|d=1)] \mid d=1\right) \\ &= E\left(\zeta(x)\cdot[x-E(x|x>\xi)] \mid x>\xi\right) \\ &= E\left\{[\zeta(x)-\zeta[E(x|x>\xi)]]\cdot[x-E(x|x>\xi)] \mid x>\xi\right\} \\ &\geq 0 \end{aligned}$$

by mean-monotonicity, since the final expectation is an integral of non-negative terms.

Corollaries 8 and 9 follow from similar arguments.

3. Estimation and Related Questions

We now discuss some of the salient aspects about estimation when there is a censored regressor in an IV context. We begin with complete case analysis, and the closely related issues of using censored instruments. We then discuss methods of estimation that makes use of the whole sample. Our discussion begins with the bivariate case as before, and we then generalize to the multivariate case.

3.1. Complete Case Analysis

One approach to estimation is to ignore the information in observations that are censored. That is, one can do "complete case analysis," by dropping any observations that are censored prior to estimation. In the case of OLS regression with censored regressors, complete case analysis permits consistent estimation, as noted by Rigobon and Stoker (2005). We now discuss whether the same approach applies to the IV situation. The answer is generally yes, provided that stronger conditions are obeyed by the instrument.

In the bivariate case that we have been discussion, dropping the censored observations amounts to computing

$$\hat{c}^* = \frac{\sum_{i=1,d_i=0}^{n} (z_i - \bar{z})(y_i - \bar{y})}{\sum_{i=1,d_i=0}^{n} (z_i - \bar{z})(x_i - \bar{x})} \rightarrow \frac{Cov(z, y|d=0)}{Cov(z, x|d=0)} = \beta + \frac{Cov(z, \varepsilon|d=0)}{Cov(z, x|d=0)} \tag{40}$$

We continue to assume $Cov(z, x|d=0) \neq 0$.

We have that

$$0 = Cov(z, \varepsilon) = (1-p)\, Cov(z, \varepsilon|d=0) + pCov(z, \varepsilon|d=1). \tag{41}$$

Therefore, a necessary condition for no bias in complete case analysis is that either $Cov(z, \varepsilon|d=0) = 0$ or $Cov(z, \varepsilon|d=1) = 0$. That is, the censoring must not eliminate (z, ε) values which are correlated.

A sufficient condition for this structure is when ε is mean independent of d and z; as in

$$E(\varepsilon|d=d_0, z=z_0) = 0 \tag{42}$$

This implies $E(\varepsilon|d=1) = 0$ and $E(z\varepsilon|d=1) = E(dz\varepsilon)/p = 0$, so that $Cov(z, \varepsilon|d=1) = 0$. In sum, the IV estimator computed using complete cases will be consistent provided that censoring doesn't distort the basic structure between instrument and disturbance. Clearly z must be a valid instrument for the complete cases, which is assured by (42), a stronger condition than $Cov(z, \varepsilon) = 0$.

There is nothing in this argument that is specific to the bivariate case, as the same remarks apply to the multivariate case. After a brief hiatus on censoring of instruments, we consider estimation with the full sample.

3.2. Censored Instruments

We now consider a related question, with virtually the same answer. Suppose instead of a censored regressor, we have a censored instrument. Does this introduce inconsistency in the IV

estimator? In general the answer is no, with stronger conditions on the original instrument.

We again assume the bivariate framework (1), (2), but instead of observing the instrument z_i, we observe a censored version

$$z_i^{cen} = (1 - d_i) z_i + d_i \zeta \tag{43}$$

where d_i is a general censoring indicator as before. Suppose that we use z_i^{cen} to instrument x_i, forming

$$\tilde{c} = \frac{\sum_{i=1}^n (z_i^{cen} - \bar{z}^{cen})(y_i - \bar{y})}{\sum_{i=1}^n (z_i^{cen} - \bar{z}^{cen})(x_i - \bar{x})} \to \frac{Cov(z^{cen}, y)}{Cov(z^{cen}, x)} = \beta + \frac{Cov(z^{cen}, \varepsilon)}{Cov(z^{cen}, x)} \tag{44}$$

We clearly must assume that $Cov(z^{cen}, x) \neq 0$.

Since

$$0 = Cov(z, \varepsilon) = Cov(z^{cen}, \varepsilon) + Cov(d(z - \zeta), \varepsilon) \tag{45}$$

the bias term in (44) vanishes if $Cov(d(z - \zeta), \varepsilon) = 0$. In other words, there is no bias if censoring produces another valid instrument. If we observed the data without censoring, then the original instrument z is comprised of two valid instruments z^{cen} and $d(z - \zeta)$.

It is easy to verify that

$$Cov(d(z - \zeta), \varepsilon) = pCov(z, \varepsilon | d = 1) + p(E(z|d = 1) - \zeta) E(\varepsilon | d = 1) \tag{46}$$

If $E(\varepsilon|d = 1) = 0$ or censoring is to the mean $\zeta = E(z|d = 1)$, then the conditions for the validity of a censored instrument match the conditions for consistent estimation with complete cases, as discussed above. In particular, if ε is mean independent of d and z as in (42), then using the censored instrument induces no bias.

Other sufficient conditions can be found. For instance, if ε is mean independent of z, and d is determined by z as in $E(d|\ z = z_0, \varepsilon = \varepsilon_0) = f(z_0)$, then IV with the censored instrument will involve no bias. This structure exists with top-coded or bottom-coded instruments, and various versions of random censoring. Mean independence of ε of z (or of d and z) is substantially stronger than $Cov(z, \varepsilon) = 0$, although it may be valid in applications (e.g. with a natural experiment, or other kind of random assignment).

3.3. Estimation with the Full Sample

We now discuss estimation using the full sample of observations (including those that are censored). For this, we consider a multivariate model

$$y_i = \alpha + \beta x_i + \phi' w_i + \varepsilon_i \quad i = 1, ..., n \tag{47}$$

where w_i is a vector of predictors. We assume that x_i is endogenous and we have the additional instrument z_i. We take w_i as exogenous, but could easily extend our remarks to the case where components of w_i are endogenous and we have additional instruments. We observe x_i^{cen} of (3), not x_i. Coefficient estimates of

$$y_i = a + bx_i^{cen} + f^{'} w_i + \varepsilon_i \quad i = 1, ..., n \tag{48}$$

using z_i and w_i as instruments, will be biased as estimators of α, β and ϕ for the reasons we have discussed at length. We could get consistent estimates from complete cases (by dropping all censored observations), provided the instrument conditions that we discussed above are satisfied.

There are at least two reasons why it is desirable to be able to consistently estimate α, β and ϕ of (47) using the full sample of data. The first is to exploit statistical efficiency. The censoring creates a substantial bias when p is relatively large; or in cases where the complete cases are a relatively small proportion $1 - p$ of the full sample. While one doesn't observe variations in x in the censored data, one does observe variations in z and w. Therefore, to the extent that variations in z and w can provide some inference about variations x, a consistent method that uses the full sample can increase efficiency over analysis of only the complete cases..

The second reason is more pragmatic and oriented toward empirical practice. When censoring is not random, dropping the censored observations will amount to choosing a special, non-random subset of the data. If the model (47) is correctly specified and applies to all possible subsets, then the only real benefit to estimation with the full sample is enhanced efficiency, as outlined before. However, in practice linear models are used in an exploratory fashion, to see how economic effects may vary across different data segments. In this spirit, estimation with the full sample includes all kinds of observations, not just those that failed to be censored. To the extent that estimates differ between those computed from complete cases and those computed with the full sample, one can infer that the censored group displays different effects than the uncensored ones. The empirical work in Rigobon and Stoker (2005) provides a example of this (with OLS estimators). There, the wealth effects on consumption were studied, with observed wealth censored at zero (bottom-coded as a measure of net wealth position). Estimation by including the censored data amounts to adding poor households to the analysis (zero or negative net wealth), that are omitted in a complete case analysis.

At this writing, we are studying a few proposals for estimation with a full sample. As in the OLS case, the issues are subtle in terms of what the available statistical information is and how best to use it.

One approach follows Rigobon and Stoker (2005), in completing the model for the censored data, and then pooling the complete cases with the censored data model. Producing a correctly specified equation for the censored observations requires assuming sufficient structure to characterize a proxy for the unobserved x variable, namely $g_1(z_i, w_i) = E(x_i \mid z_i, w_i, d_i = 1)$. The

model for the censored observations is then

$$\begin{aligned} y_i &= \alpha + \beta E\left(x_i \mid z_i, w_i, d_i = 1\right) + \phi^{'} w_i + \upsilon_i \\ &= \alpha + \beta g_1\left(z_i, w_i\right) + \phi^{'} w_i + \upsilon_i \end{aligned} \tag{49}$$

where $E\left(\upsilon_i \mid z, w_i\right) = 0$. We can pool this regression model with the (instrumented) complete case model to obtain more efficient estimates of β and ϕ. As discussed in Rigobon and Stoker (2005), this approach requires assumptions on the censoring process, as clearly, the regression $g_1\left(z_i, w_i\right)$ will take different forms depending on whether censoring is random, or reflects top-coding, or some other process.

Another promising approach is to estimate the reverse regression, following a suggestion of Jerry Hausman. In particular, because of the endogeneity, y and x have a roughly symmetric role in (47). Consequently, suppose we interchange them, as in the the reverse regression

$$x_i = -\frac{\alpha}{\beta} + \frac{1}{\beta} y_i - \frac{\phi^{'}}{\beta} w_i - \frac{\varepsilon_i}{\beta} \quad i = 1, ..., n \tag{50}$$

What does full sample estimation amount to now? It amounts to estimation of (50) using the censored *dependent* variable x_i^{cen}. So, there are two questions. First, is there an instrument for the endogenous regressor y_i? It appears that z_i will do the job nicely, as $Cov\left(z_i, -\varepsilon/\beta\right) = 0$ and $Cov\left(z_i, y_i\right) \neq 0$. Second, can we use a semiparametric estimation method for the censored Tobit model in this case? For instance, can we replace y_i by its fitted value on z_i and w_i to get a consistent estimate of $1/\beta$? We are currently studying these kinds of methods as a solution to the bias introduced by censored endogenous regressors.[17]

4. Concluding Remarks

There is one theme that has emerged in our analysis of the inconsistency of IV estimators when a censored regressor is used. If the correlation structure of the data that is censored is at all similar to the correlation structure of the data that is not censored, then there will be expansion bias, with IV estimates too large. For that not to be the case, the censored and uncensored data need to be radically different. For instance, when the censoring is uncorrelated with the instrument, expansion bias holds unless the covariance between instrument and predictor for the censored data is of the opposite sign for that of the uncensored data. With top-coding or

[17] Among others, Blundell and Smith (1989,1994) provide parametric methods of estimation for models with censored dependent variables and endogenous regressors, and Hong and Tamer (2003) and Blundell and Powell (2004) propose semiparametric methods.

bottom-coding, expansion bias arises unless the covariance for censored data is of opposite sign and of substantial magnitude. With independent censoring, the censored and uncensored data have the same structure by construction, so expansion bias always results.

Our results are based on a bivariate model, which is certainly stylized. As in Rigobon and Stoker (2005), it is natural to expect that expansion bias will transmit across various regressors, depending upon the correlation structure. This analysis would be worthwhile to carry out, especially for situations where the endogenous regressor represents a heterogeneous treatment effect, and other controls are used in the analysis.

At this writing, we have not fully analyzed the issues regarding consistent estimation and efficiency in the IV context. Rigobon and Stoker (2005) analyze how to use censored observations to improve efficiency in estimation of a standard linear model, and as we have suggested, some of the features of their analysis ought to apply here, where the framework is more general. To understand how much ones basic IV estimates have been affected by censoring, it is very desirable to have methods that permit estimation with the full data sample.

References

[1] Black, D.A., M.C. Berger and F.A. Scott (2000), "Bounding Parameter Estimates with Nonclassical Measurement Error," *Journal of the American Statistical Association*, 95, 739-748.

[2] Blundell, R.J. and J.L. Powell (2004), "Censored Regression Quantiles with Endogenous Regressors," Working Paper, Department of Economics, University of California at Berkeley, November.

[3] Blundell, R.W. and R. J. Smith (1989), "Estimation in a Class of Simultaneous Equation Limited Dependent Variable Models," *Review of Economic Studies*, 56, 37-58.

[4] Blundell, R.W. and R. J. Smith (1994), "Coherency and Estimation in Simultaneous Equation Models with Censored or Qualitative Dependent Variables," *Journal of Econometrics*, 64, 355-373.

[5] Davidson, R. and J. D. McKinnon (2004), *Econometric Theory and Methods*, Oxford University Press, New York.

[6] Hong, H and E. Tamer (2003), "Inference in Censored Models with Endogenous Regressors," *Econometrica*, 71, 905-932.

[7] Manski, C.F. and E. Tamer (2002) "Inference on Regressions with Interval Data on a Regressor or Outcome," *Econometrica*, 70, 519-546.

[8] Rigobon, R. and T. M. Stoker (2005), "Censored Regressors and Expansion Bias," MIT Sloan School of Management Working Paper, March 2005.

[9] Wald, A. (1940), "The Fitting of Straight Lines if Both Variables are Subject to Error," *Annals of Mathematical Statistics*, 11, 284-300.

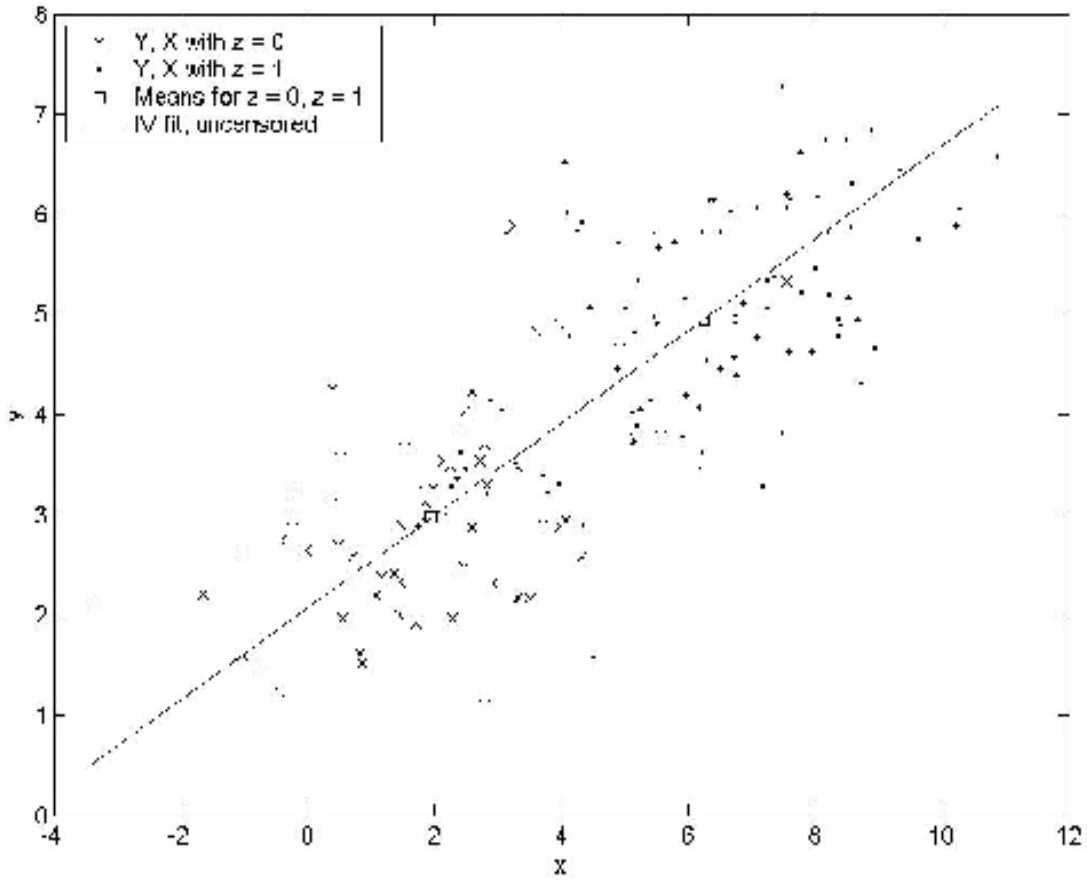

Figure 1: Uncensored Data with Random Assignment

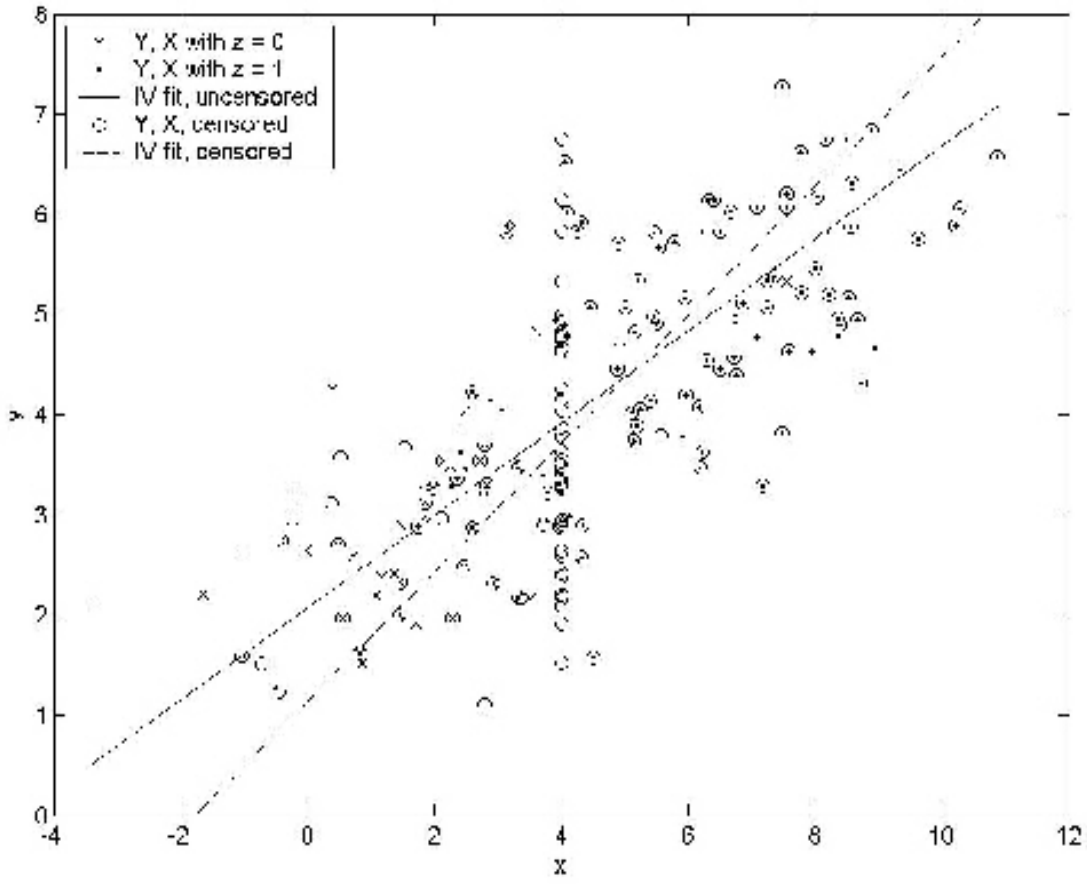

Figure 2: IV Bias with Random Censoring

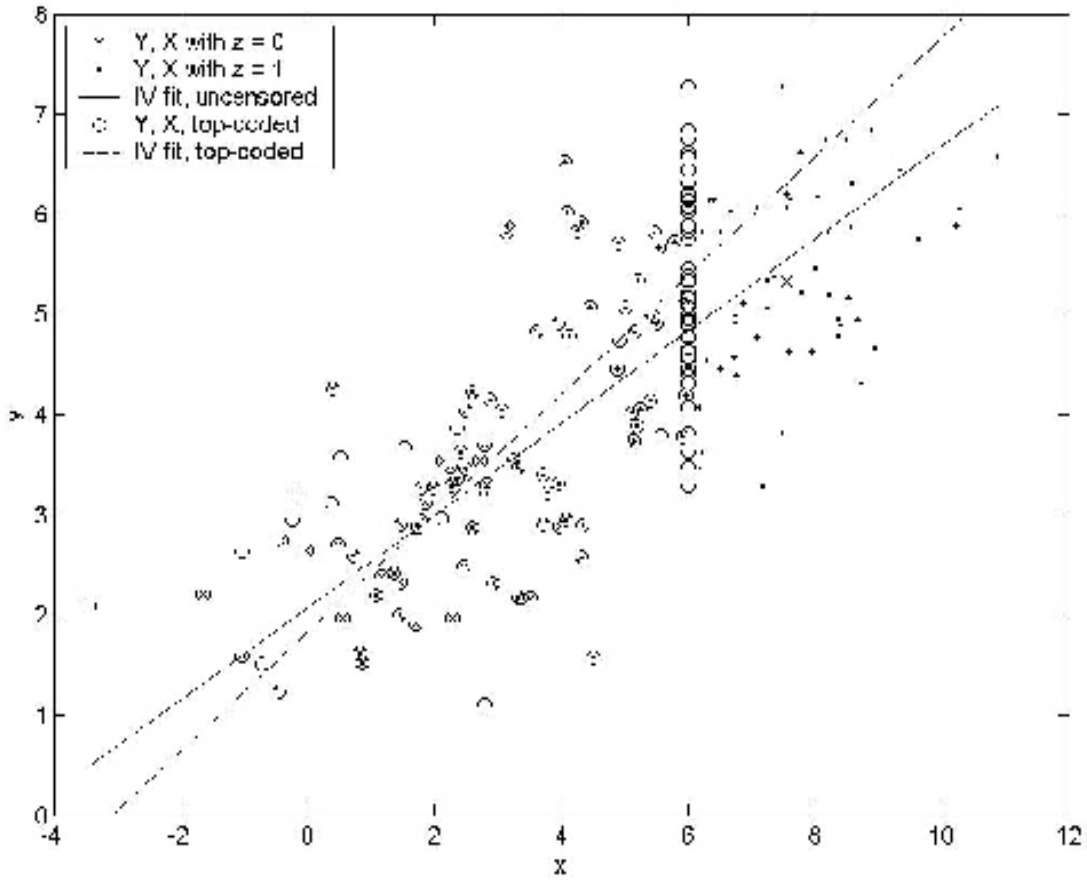

Figure 3: IV Bias with Top-Coding